

PATENT ABSTRACTS OF JAPAN

(11)Publication number : 2001-155020

(43)Date of publication of application : 08.06.2001

(51)Int.Cl.

G06F 17/30

(21)Application number : 11-334597

(71)Applicant : TOSHIBA CORP
TOSHIBA COMPUT ENG CORP

(22)Date of filing : 25.11.1999

(72)Inventor : KOBAYASHI TSUTOMU
NAKAZATO SHIGEMI
SAITO HIROMI
NISHINA TAKUYA
NAKAMOTO YUKIO
YAMAZAKI HIROSHI
MATSUKUMA TAKESHI

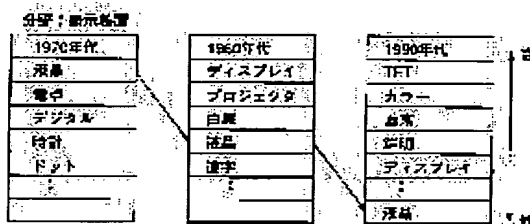
(54) DEVICE AND METHOD FOR RETRIEVING SIMILAR DOCUMENT AND RECORDING MEDIUM

(57)Abstract:

PROBLEM TO BE SOLVED: To highly accurately retrieve a similar document while considering a category, to which a document belongs, or timing of preparation.

SOLUTION: In the similar document retrieving device for retrieving a document similar to a document applied as a retrieval key out of plural retrieval object documents, the importance of a word by times is provided by weighing the appearance frequency of a word for each document corresponding to the time element of each document while considering the change of the word, with which the contents of the document are characterized, with the passage of time and similarity is calculated while considering that importance of the word. Further, since the change of the feature word with the passage of time is different corresponding to the category, to which the document belongs, as well, respective documents are sorted by categories, the importance of a word by categories and by times is provided by performing weighing corresponding to the time element

by categories and similarity is calculated while considering that importance of the word. Thus, the similar document can be highly accurately retrieved while exactly reflecting the importance of the word on the calculation of similarity.



LEGAL STATUS

[Date of request for examination]

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than

the examiner's decision of rejection or
application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision
of rejection]

[Date of requesting appeal against examiner's
decision of rejection]

[Date of extinction of right]

(19) 日本国特許庁 (J P)

(12) 公開特許公報 (A)

(11) 特許出願公開番号

特開2001-155020

(P2001-155020A)

(43) 公開日 平成13年6月8日 (2001.6.8)

(51) Int.Cl.⁷

G 0 6 F 17/30

識別記号

F I

G 0 6 F 15/40

15/403

テームト* (参考)

3 7 0 A 5 B 0 7 5

3 4 0 B

3 5 0 C

審査請求 未請求 請求項の数 8 O L (全 14 頁)

(21) 出願番号 特願平11-334597

(22) 出願日 平成11年11月25日 (1999. 11. 25)

(71) 出願人 000003078

株式会社東芝

神奈川県川崎市幸区堀川町72番地

(71) 出願人 000221052

東芝コンピュータエンジニアリング株式会
社

東京都青梅市新町3丁目3番地の1

(72) 発明者 小林 勉

東京都青梅市末広町2丁目9番地 株式会
社東芝青梅工場内

(74) 代理人 100058479

弁理士 鈴江 武彦 (外6名)

最終頁に続く

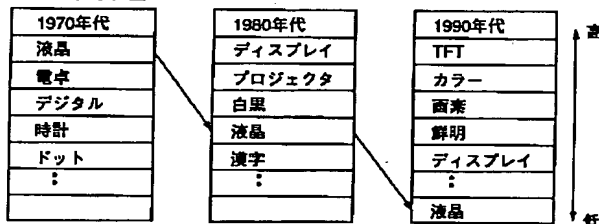
(54) 【発明の名称】 類似文書検索装置、類似文書検索方法及び記録媒体

(57) 【要約】

【課題】 文書が属する分野や作成時期を考慮して類似文書を高精度に検索する。

【解決手段】 複数の検索対象文書の中から検索キーとして与えられた文書と類似する文書を検索する類似文書検索装置において、文書の内容を特徴付ける単語の時間的な変化に着目し、各文書毎の単語の出現頻度に各文書の時間的な要素に応じた重み付けを行うことで、時間別の単語の重要度を得て、その単語の重要度を加味した類似度計算を行う。さらに、特徴単語の時間的な変化は文書が属する分野によっても異なるため、各文書を分野別に分類し、その分野別の時間的な要素に応じた重み付けを行うことで、分野別かつ時間別の単語の重要度を得て、その単語の重要度を加味した類似度計算を行う。これにより、類似度計算に単語の重要度を的確に反映させて、高精度に類似文書を検索することができる。

分野：表示装置



【特許請求の範囲】

【請求項1】 複数の検索対象文書の中から検索キーとして与えられた文書と類似する文書を検索する類似文書検索装置において、

上記各検索対象文書および上記検索キー文書のそれぞれに含まれる単語の出現頻度を求める出現頻度算出手段と、

この出現頻度算出手段によって得られた上記各検索対象文書および上記検索対象文書に対応した単語の出現頻度に、それぞれの文書の時間的な要素に応じた重み付けを行う単語重み付け手段と、

この単語重み付け手段によって重み付けされた単語出現頻度に基づいて上記各検索対象文書と上記検索キー文書との類似度を算出する類似度算出手段と、

この類似度算出手段によって得られた類似度に基づいて上記各検索対象文書の中から上記検索対象文書と類似する文書を検索結果として出力する出力手段とを具備したことを特徴とする類似文書検索装置。

【請求項2】 上記単語重み付け手段は、上記出現頻度算出手段によって得られた上記各検索対象文書および上記検索対象文書に対応した単語の出現頻度に、それぞれの文書が属する分野毎の時間的な要素に応じて重み付けを行うことを特徴とする請求項1記載の類似文書検索装置。

【請求項3】 複数の検索対象文書の中から検索キーとして与えられた文書と類似する文書を検索する類似文書検索装置において、

上記各検索対象文書を時間的要素に基づいて分類する分類手段と、

この分類手段による分類別に上記各検索対象文書に含まれる単語の出現頻度を求める第1の出現頻度算出手段と、

上記各検索対象文書および上記検索キー文書のそれぞれに含まれる単語の出現頻度を求める第2の出現頻度算出手段と、

上記第1の出現頻度算出手段によって得られた分類別の単語出現頻度に基づいて、上記第2の出現頻度算出手段によって得られた上記各検索対象文書および上記検索キー文書のそれぞれの単語の出現頻度に重み付けを行う単語重み付け手段と、

この単語重み付け手段によって重み付けされた単語出現頻度に基づいて上記各検索対象文書と上記検索キー文書との類似度を算出する類似度算出手段と、

この類似度算出手段によって得られた類似度に基づいて上記各検索対象文書の中から上記検索対象文書と類似する文書を検索結果として出力する出力手段とを具備したことを特徴とする類似文書検索装置。

【請求項4】 上記分類手段は、上記各検索対象文書をそれぞれが属する分野毎の時間的要素に基づいて分類することを特徴とする請求項3記載の類似文書検索装置。

【請求項5】 複数の検索対象文書の中から検索キーとして与えられた文書と類似する文書を検索する類似文書検索方法において、

上記各検索対象文書および上記検索キー文書のそれぞれに含まれる単語の出現頻度を求め、

上記各検索対象文書および上記検索対象文書に対応した単語の出現頻度に、それぞれの文書の時間的要素に応じた重み付けを行い、

この重み付け後の単語出現頻度に基づいて上記各検索対象文書と上記検索キー文書との類似度を算出し、

この算出された類似度に基づいて上記各検索対象文書の中から上記検索対象文書と類似する文書を検索結果として出力することを特徴とする類似文書検索方法。

【請求項6】 複数の検索対象文書の中から検索キーとして与えられた文書と類似する文書を検索する類似文書検索方法において、

上記各検索対象文書および上記検索キー文書のそれぞれに含まれる単語の出現頻度を求め、

上記各検索対象文書および上記検索対象文書に対応した単語の出現頻度に、それぞれの文書が属する分野毎の時間的要素に応じた重み付けを行い、

この重み付け後の単語出現頻度に基づいて上記各検索対象文書と上記検索キー文書との類似度を算出し、

この算出された類似度に基づいて上記各検索対象文書の中から上記検索対象文書と類似する文書を検索結果として出力することを特徴とする類似文書検索方法。

【請求項7】 複数の検索対象文書の中から検索キーとして与えられた文書と類似する文書を検索する類似文書検索機能を備えたコンピュータに、

上記各検索対象文書および上記検索キー文書のそれぞれに含まれる単語の出現頻度を求める機能と、

上記各検索対象文書および上記検索対象文書に対応した単語の出現頻度に、それぞれの文書の時間的要素に応じた重み付けを行う機能と、

この重み付け後の単語出現頻度に基づいて上記各検索対象文書と上記検索キー文書との類似度を算出する機能と、

この算出された類似度に基づいて上記各検索対象文書の中から上記検索対象文書と類似する文書を検索結果として出力する機能とを実行させるプログラムを記録したコンピュータ読取り可能な記録媒体。

【請求項8】 複数の検索対象文書の中から検索キーとして与えられた文書と類似する文書を検索する類似文書検索機能を備えたコンピュータに、

上記各検索対象文書および上記検索キー文書のそれぞれに含まれる単語の出現頻度を求める機能と、

上記各検索対象文書および上記検索対象文書に対応した単語の出現頻度に、それぞれの文書が属する分野毎の時間的要素に応じた重み付けを行う機能と、

この重み付け後の単語出現頻度に基づいて上記各検索対

象文書と上記検索キー文書との類似度を算出する機能と、

この算出された類似度に基づいて上記各検索対象文書の中から上記検索対象文書と類似する文書を検索結果として出力する機能とを実行させるプログラムを記録したコンピュータ読取り可能な記録媒体。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】本発明は、データベースに登録された複数の検索対象文書の中から類似文書と類似する文書を検索する類似文書検索装置であって、特に各文書に含まれる単語の重要度を考慮して類似文書の検索を行う類似文書検索装置と、この装置に用いられる類似文書検索方法及び記録媒体に関する。

【0002】

【従来の技術】従来、例えば引用文献等として用いられる各種文書をデータベース化しておき、その中から指定された文書（以下、検索キー文書と称す）に類似する文書を自動検索するシステムがある。このようなシステムでは、検索キー文書に含まれている単語と検索対象となる文書に含まれている単語とを比較し、共通する単語の種類、出現場所、出現回数などからベクトル空間法などにより類似度を算出して、類似度の高い文書を検索結果として出力している。

【0003】このような類似文書検索では、検索キー文書や検索対象となる文書から、その文書の内容を特徴付ける単語を抽出することが重要となる。通常、各文書に含まれる単語の出現頻度を単語種毎に求め、出現頻度の低い単語をその文書の内容を特徴付ける単語として判断している。

【0004】

【発明が解決しようとする課題】上述したように、従来、単語の出現頻度によって特徴単語を判断していた。しかしながら、検索対象となる文書が膨大にあり、しかも、長い年月に亘って蓄積された場合には、単語の出現頻度からだけでは特徴単語を的確に判断することはできない。これは、各文書が属する分野によって、その文書の特徴付ける単語の種類が異なり、また、同じ分野の中でも文書の作成時期（発行時期）によって特徴単語が異なるからである。

【0005】本発明は上記のような点に鑑みなされたもので、文書が属する分野や作成時期を考慮して類似文書を高精度に検索することのできる類似文書検索装置、類似文書検索方法及び記録媒体を提供することを目的とする。

【0006】

【課題を解決するための手段】本発明の類似文書検索装置は、文書の特徴付ける単語の時間的な変化に着目し、各文書に含まれる単語の出現頻度に各文書の時間的な要素に応じた重み付けを行うようにしたものである。

【0007】具体的には、各検索対象文書および検索キー文書のそれぞれに含まれる単語の出現頻度を求める出現頻度算出手段と、この出現頻度算出手段によって得られた上記各検索対象文書および上記検索対象文書に対応した単語の出現頻度に、それぞれの文書の時間的な要素に応じた重み付けを行う単語重み付け手段とを備え、この単語重み付け手段による重み付け後の単語出現頻度に基づいて上記各検索対象文書と上記検索キー文書との類似度を算出し、その類似度に基づいて上記各検索対象文書の中から上記検索対象文書と類似する文書を検索結果として出力するようにしたものである。

【0008】このような構成によれば、各文書毎の単語の出現頻度にそれぞれの文書の時間的な要素に応じた重み付けを行うことで、時間別の単語の重要度を得ることができ、その単語の重要度を加味した類似度計算により、検索キー文書と類似する文書を的確に検索することができる。

【0009】さらに、本発明は、上述した特徴単語の時間的な変化は分野によっても異なることに着目し、各検索対象文書および検索対象文書に対応した単語の出現頻度に、それぞれの文書が属する分野毎の時間的な要素に応じて重み付けを行うことを特徴とする。

【0010】このように、分野毎の時間的な要素に応じた重み付けを行うことで、分野別かつ時間別の単語の重要度を得ることができ、その単語の重要度を加味した類似度計算により、検索キー文書と類似する文書をさらに的確に検索することができる。

【0011】

【発明の実施の形態】まず、本発明の実施形態を説明する前に理解を容易にするため、本発明の類似文書検索装置の概要について説明する。

【0012】本発明の類似文書検索装置は、検索キーとして与えられた文書と類似する文書を複数の検索対象文書の中から検索するものである。このような類似文書の検索では、文書の内容を特徴付ける単語を抽出することが重要となる。

【0013】ここで、図1に示すように、文書の内容を特徴付ける単語は時代の流れによって異なる。例えば、「表示装置」といった分野（大分類）では、「液晶」といった単語は1970年代では重要度の高い単語として扱われていた。しかし、技術の発展に伴い、「液晶」といった単語は一般化し、1990年代ではその重要度は低く、代わって別の単語（TFT等）の重要度が高くなる。このように、文書の特徴付ける単語は時間的に変化する。したがって、従来のように単に単語の出現頻度だけから検索キー文書との類似度を求めていた方法では、単語の重要度を正確に反映させることはできない。

【0014】そこで、本発明では、このような特徴単語の時間的な変化に着目し、各文書毎の単語の出現頻度に各文書の時間的な要素に応じた重み付けを行うことで、

時間別の単語の重要度を得て、その単語の重要度を加味した類似度計算を行うようにしたものである。

【0015】また、このような特徴単語の時間的な変化は、文書が属する分野によっても異なる。例えば、コンピュータ分野などでは、他の分野よりも時間的な変化が速い。そこで、各文書を分野別に分類し、その分野別の時間的な要素に応じた重み付けを行うことで、分野別かつ時間別の単語の重要度を得て、その単語の重要度を加味した類似度計算を行うようにしたものである。

【0016】以下に、このような類似文書の検索を実現するための実施形態について説明する。

【0017】図2は本発明の類似文書検索装置の構成を示すブロック図である。なお、本装置は、例えば磁気ディスク等の記録媒体に記録されたプログラムを読み込み、このプログラムによって動作が制御されるコンピュータによって実現される。

【0018】図2に示すように、本装置は制御装置1、入力装置2、表示装置3、外部記憶装置4から構成される。制御装置1は、制御部1a（CPU）とメモリ部1bからなり、所定のプログラムに従って本装置全体の制御を行う。入力装置2は、例えばキーボードやマウスなどからなり、データの入力や指示を行う場合に用いられる。表示装置3は、例えばCRT（Cathode-ray tube）やLCD（Liquid Crystal Display）などからなり、類似検索結果などを表示する。外部記憶装置4は、例えば磁気ディスク装置や光ディスク装置などからなり、本装置で扱われる各種データを保持する。

【0019】この外部記憶装置4には、検索対象となる各文書のデータ（テキスト情報）、類似検索に必要な各文書の単語情報の他、各種データファイルF1～F6が設けられている。F1は検索文書大分類識別子データファイル、F2は検索文書大分類データファイル、F3は検索文書時間区分識別子データファイル、F4は検索文書時間区分データファイルである。また、F5は全検索文書大分類情報データファイル、F6は全検索文書時間区分データファイルである。

【0020】図3は制御装置1の内部構成を示した図である。

【0021】制御装置1は制御部1aとメモリ部1bからなっている。制御部1aは、後述するデータベース作成処理や類似文書検索処理を実行するための処理部101～126から構成される。また、メモリ部1bはこれらの処理部101～126に必要なデータを格納するためのバッファ201～219から構成される。

【0022】初期化部101は、上記各バッファ201～219の初期化を行う。

【0023】検索文書大分類データ読み込み部102は、外部記憶装置4に格納されている検索文書大分類識別子データファイルF1の内容を検索文書大分類識別子格納バッファ201に格納する。この検索文書大分類

識別子データファイルF1は、図10に示すように、大分類（分野別）の識別子を示したファイルである。

【0024】検索文書大分類データ読み込み部103は、外部記憶装置4に格納されている検索文書大分類データファイルF2の内容を検索文書大分類格納バッファ202に格納する。この検索文書大分類データファイルF2は、図11に示すように、大分類IDと大分類データとを対応付けたファイルである。

【0025】検索文書時間区分識別データ読み込み部104は、外部記憶装置4に格納されている検索文書時間区分識別子データファイルF3の内容を検索文書時間区分識別子格納バッファ203に格納する。この検索文書時間区分識別子データファイルF3は、図12に示すように、時間区分別の識別子を示したファイルである。

【0026】検索文書時間区分データ読み込み部105は、外部記憶装置4に格納されている検索文書時間区分データファイルF4の内容を検索文書時間区分格納バッファ204に格納する。この検索文書時間区分データファイルF4は、図13に示すように、時間区分IDと時間区分とを対応付けたファイルである。

【0027】検索文書読み込み部106は、外部記憶装置4に格納されている検索文書（検索対象となる文書）に関する情報をデータベース化するために、各文書のテキスト情報を外部記憶装置4から順に読み込み、検索文書格納バッファ205に格納する。

【0028】検索文書大分類分け部107は、検索文書格納バッファ205に格納された検索文書を大分類（分野）毎に分類するためにものであり、当該検索文書の内容から検索文書大分類識別子および大分類データを抽出し、検索文書大分類格納バッファ202を参照して、当該検索文書の大分類データとその大分データに対応する大分類IDを検索文書大分類情報格納バッファ206に格納する。さらに、当該検索文書に対する検索文書IDと検索文書大分類情報格納バッファ206に格納されている大分類IDを全検索文書大分類情報格納バッファ208に格納する。

【0029】全検索文書大分類情報書き込み部108は、全検索文書大分類情報格納バッファ208に格納されている全検索文書の大分類情報を外部記憶装置4に格納する。

【0030】検索文書時間区分識別子抽出部109は、検索文書格納バッファ205に格納されている検索文書を時間区別別に分類するためのものであり、当該検索文書から検索文書時間区分識別子および時間区分を抽出し、検索文書時間区分データ格納バッファ204を参照して、当該検索文書の時間区分とその時間区分に対応する時間区分IDを検索文書時間区分格納バッファ207に格納する。さらに、当該検索文書に対する検索文書IDと検索文書時間区分格納バッファ207に格納されている時間区分IDを全検索文書時間区分格納バッファ2

09に格納する。

【0031】全検索文書時間区分書き込み部110は、全検索文書時間区分対応格納バッファ209に格納されている全検索文書の時間区分を外部記憶装置4に格納する。

【0032】検索文書単語抽出部111は、検索文書格納バッファ205に格納されている検索文書から単語の切り出しを行う。そして、切り出した各単語の中からその文書の内容を表す上でキーとなる単語を抽出し、その単語種毎に検索文書単語情報格納バッファ210に格納する。単語の切り出しは、形態素解析などにより行い、その文書の内容を表す上でキーとなる単語の単語種は品詞情報（例えば「名詞」や「サ変名詞」）を使って表現する。

【0033】検索文書単語出現頻度算出部112は、検索文書単語抽出部110により抽出された個々のキー単語について、検索文書中での出現頻度を算出し、これを検索文書の単語情報として検索文書単語情報格納バッファ210に格納する。

【0034】検索文書単語情報書き込み部113は、検索文書単語情報格納バッファ210に格納されている検索文書の単語情報を外部記憶装置4に格納する。

【0035】全検索文書大分類情報読み込み部114は、外部記憶装置4に格納されている検索文書毎の大分類情報（全検索文書大分類情報データファイルF5）を全検索文書大分類情報対応格納バッファ208に格納する。

【0036】全検索文書時間区分読み込み部115は、外部記憶装置4に格納されている検索文書毎の時間区分（全検索文書時間区分データファイルF6）を全検索文書時間区分対応格納バッファ209に格納する。

【0037】検索文書単語情報読み込み部116は、外部記憶装置4に格納されている検索文書単語情報（単語の出現頻度情報）を検索文書毎に呼び出し、検索文書単語情報格納バッファ210に格納する。

【0038】検索文書単語出現頻度集計部117は、大分類および時間区分別に各単語の出現頻度を算出し、検索文書単語頻度格納バッファ211に格納する。詳しくは、検索文書単語情報格納バッファ210に読み込まれた検索文書単語情報に対して、検索文書大分類格納バッファ202と全検索文書大分類情報格納バッファ208を参照して、検索文書大分類格納バッファ202に格納された大分類ID別に各単語の出現頻度を集計する。また、検索文書時間区分データ格納バッファ204と検索文書時間区分対応格納バッファ209を参照し、検索文書情報区分データ格納バッファ204に格納された時間区分ID別に各単語の出現頻度を集計する。

【0039】検索文書単語出現頻度書き込み部118は、検索文書単語頻度格納バッファ211に格納されている大分類別時間区分別の単語出現頻度情報を外部記憶

装置4に格納する。

【0040】検索キー文書読み込み部119は、入力装置2から入力された検索キー文書のテキスト情報を検索キー文書格納バッファ212に格納する。

【0041】検索キー単語抽出部120は、検索キー文書格納バッファ212に格納されている検索キー文書から単語の切り出しを行う。そして、切り出した各単語の中からその文書の内容を表す上でキーとなる単語を抽出し、その単語種を検索キー文書単語情報格納バッファ213に格納する。上記検索文書単語抽出部111と同様に、単語の切り出しは形態素解析などにより行い、その文書の内容を表す上でキーとなる単語の単語種は品詞情報（例えば「名詞」や「サ変名詞」）を使って表現する。

【0042】検索キー単語出現頻度算出部121は、検索キー単語抽出部119により抽出された個々のキー単語について、検索キー文書中での出現頻度を算出し、これを検索キー文書の単語情報として検索キー文書単語情報格納バッファ213に格納する。

【0043】検索文書単語頻度読み込み部122は、外部記憶装置4に格納されている検索文書単語出現頻度情報を大分類別時間区分別の検索文書単語頻度格納バッファ211に格納する。

【0044】ノルム情報算出部123は、検索文書単語情報あるいは検索キー文書単語情報に格納された単語ID毎の単語出現頻度を要素とする1次元ベクトルのノルムを計算する。なお、ノルムとは、ベクトルの大きさを表わすものである。その際、検索文書単語頻度格納バッファ211に格納されている大分類別時間区分別の単語頻度情報を参照し、該当する単語に対する頻度情報を加重してノルム計算を行い、その結果を検索文書ノルム情報格納バッファ214あるいは検索キー文書ノルム情報格納バッファ215に格納する。

【0045】共通単語抽出部124は、検索キー文書単語情報格納バッファ213に格納されている検索キー文書の単語情報と検索文書単語情報格納バッファ210に格納されている検索文書の単語情報とを比較して、両者で共通する単語の種類とその出現頻度情報を共通単語情報格納バッファ216に格納する。

【0046】類似度算出部125は、共通単語情報格納バッファ216に格納されている共通単語に関する情報に基づき検索キーと検索文書との類似度を算出し、その類似度値を類似度格納バッファ217に格納する。

【0047】検索結果出力部126は、類似度格納バッファ217に格納されている各検索文書に対する類似度値を高い順に並べ替えて、検索結果出力バッファ218に格納し、さらに、検索結果出力バッファ218の内容を表示装置3に出力する。

【0048】次に、本装置の動作について説明する。

【0049】なお、ここでは、(a) データベース作成

処理と、(b)類似文書作成処理に分けて、それぞれの処理動作について説明する。

【0050】(a)データベース作成処理

まず、本装置にて実行されるデータベース作成処理について説明する。

【0051】図4はデータベース作成処理の動作を示すフローチャートである。

【0052】検索文書データベースを作成する手順として、検索対象となる各文書を1件ずつ処理して、各文書の単語情報を作成し(ステップA11)、全ての文書の処理が終了したら、上記単語情報を使って単語頻度に関する情報を作成する(ステップA12)。ステップA11における検索文書登録処理の詳細を図5、ステップA12における検索文書単語出現頻度作成処理の詳細を図6に示す。

【0053】図5は上記図4のステップA11における検索文書登録処理の動作を示すフローチャートである。

【0054】制御装置1は、まず、制御部1aの初期化部101を起動し、メモリ部1bの全バッファ部の初期化を行う(ステップB11)。また、各データ読み込み部102~105を起動して、外部記憶装置4の検索文書大分類識別子データファイルF1、検索文書大分類データファイルF2、検索文書時間区分識別子データファイルF3、検索文書時間区分データファイルF4の情報を読み込み、それぞれ検索文書大分類識別子格納バッファ201、検索文書大分類格納バッファ202、検索文書時間区分識別子データ格納バッファ203、検索文書時間区分データ格納バッファ204に格納する(ステップB12)。

【0055】図15に検索文書大分類識別子格納バッファ201のデータ格納例、図16に検索文書大分類格納バッファ202のデータ格納例、図17に検索文書時間区分識別子データ格納バッファ203のデータ格納例、図18に検索文書時間区分データ格納バッファ204のデータ格納例を示す。

【0056】ここで、外部記憶装置4にデータベースに登録する文書がある場合には(ステップB13のYes)、検索文書読み込み部106が外部記憶装置4からその文書のテキスト文書を読み込み、これを検索対象として検索文書格納バッファ205に格納する(ステップB14)。具体例として、例えば図14に示すような内容を有するテキスト文書を検索文書の1つとして検索文書格納バッファ205に格納したとする。

【0057】検索文書の読み込み後、制御装置1は検索文書格納バッファ205に格納された検索文書を先頭から順に走査していく(ステップB15)。その際、当該検索文書から大分類情報を抽出できたか否かチェックする(ステップB16)、まだならば(ステップB16のNo)、検索文書大分類分け部107を起動し、検索文書大分類情報識別子格納バッファ201に格納されてい

る識別子情報を参照して当該文書から大分類情報を抽出する(ステップB17)。

【0058】具体的に説明すると、例えば図14に示すようなテキスト文書であれば、検索文書大分類情報識別子格納バッファ201に格納されている識別子情報

(「[大分類] ???」)を参照し、1行目の「[大分類] A01」から識別子情報の「???」に相当する

「A01」を大分類情報として抽出し、この大分類情報「A01」に対応する大分類ID「1」と共に検索文書大分類情報格納バッファ206に格納する。図19にこのときのデータ格納例を示す。その際、当該検索文書のID「1」と検索文書大分類情報格納バッファ206に格納されている大分類ID「1」とを対にして全検索文書大分類情報格納バッファ208に格納する。図20にこのときのデータ格納例を示す。

【0059】続いて、制御装置1は検索文書格納バッファ205を参照し、文書終端でなければ(ステップB15のNo)、時間区分を抽出できたか否かをチェックする(ステップB18)。まだならば(ステップB18のNo)、検索文書時間区分識別子抽出部109を起動し、検索文書時間区分識別子データ格納バッファ203に格納されている識別子情報を参照して当該検索文書から時間区分を抽出する(ステップB19)。

【0060】具体的に説明すると、例えば図14に示すようなテキスト文書であれば、検索文書時間区分識別子データ格納バッファ203に格納されている識別子情報

(「[出願日] ???年」)を参照し、1行目の「[出願日] 1999年(平成5年)9月27日」から識別子情報の「????」に相当する「1999」を時間区分として抽出し、この時間区分「1999」に対応する時間区分ID「1」と共に検索文書時間区分格納バッファ207に格納する。図21にこのときのデータ格納例を示す。その際、当該検索文書に対する検索文書ID「1」と検索文書時間区分格納バッファ207に格納されている時間区分ID「1」とを対にして全検索文書時間区分格納バッファ209に格納する。図22にこのときのデータ格納例を示す。

【0061】また、大分類情報抽出済みであり(ステップB16のYes)、かつ、時間区分抽出済みでならば(ステップB18のYes)、制御装置1は検索対象単語抽出部111を起動し、検索文書格納バッファ205に格納された検索文書のテキスト情報を形態素解析して単語の切り出しを行い、その切り出した各単語の中から文書内容を表すキー単語を抽出し、そのキー単語の単語種(例えば品詞情報)を検索文書単語情報格納バッファ210に格納する(ステップB20)。

【0062】続いて、検索文書単語出現頻度算出部112を起動し、当該検索文書全体での単語の出現頻度を算出し、その結果を検索文書単語情報格納バッファ210に格納する(ステップB21)。図23にこのときのデ

ータ格納例を示す。この場合、単語と頻度が対について格納される。例えば、キー単語「文書」が当該検索文書の中で2回出現している場合は、頻度として「2」が格納される。

【0063】文書の終端まで達したら（ステップB15のYes）、検索文書単語情報書き込み部113を起動し、検索文書単語情報格納バッファ210に格納された情報を検索文書の単語情報として外部記憶装置4に格納、蓄積する（ステップB22）。また、全検索文書大分類情報書き込み部108を起動し、全検索文書大分類情報格納バッファ208に格納された全検索文書の大分類情報を外部記憶装置4に格納し、全検索文書時間区分書き込み部110を起動し、全検索文書時間区分対応格納バッファ209に格納された全検索文書の時間区分を外部記憶装置4に格納する。

【0064】以上のような処理を登録対象となる検索文書のすべてについて、繰り返し実行する。

【0065】次に、検索文書単語出現頻度作成処理について説明する。

【0066】図6は上記ステップA12における検索文書単語出現頻度作成処理の動作を示すフローチャートである。

【0067】まず、制御装置1は全検索文書大分類情報読み込み部114を起動し、外部記憶装置4から文書対大分類の情報（全検索文書大分類情報データファイルF5）を読み込んで全検索文書大分類情報格納バッファ208に格納すると共に、全検索文書時間区分読み込み部115を起動し、外部記憶装置4から文書対時間区分の情報（全検索文書時間区分データファイルF6）を読み込んで全検索文書時間区分格納バッファ209に格納する（ステップC11）。

【0068】そして、制御装置1は全検索文書時間区分格納バッファ209に格納された検索文書IDと時間区分IDを参照して、全文書に対し、以下のような集計処理を行う（ステップC12）。

【0069】すなわち、検索文書単語情報読み込み部116によって、該当する検索文書IDに対応する検索文書単語情報を外部記憶装置4から読み込み、これを検索文書単語情報格納バッファ210に読み込む（ステップC13）。

【0070】続いて、検索文書単語出現頻度算出部117によって、検索文書単語情報格納バッファ210に読み込まれた当該検索文書の検索文書単語情報に関し、その文書が属する大分類別、かつ、時間区分別に全ての単語について、出現の有無を累積し、大分類別時間区分別の検索文書単語頻度格納バッファ211に格納する（ステップC14）。上記出現の有無を累積するとは、ある文書にある単語が複数出現していても、それを1回と数えるということである。

【0071】大分類別かつ時間区分別に、全ての文書に

ついて上記ステップC14の処理を行うと、検索文書単語頻度格納バッファ211の内容は図25のようになる。例えば、「大分類ID:1」の「時間区分ID:1」において「表示」という単語は120回出現し、「区分」という単語は9回出現しているという情報が格納されている。したがって、この「時間区分ID:1」に属する文書数がnとした場合、「表示」という単語は文書数nの中で120文書に出現しているということになる。

【0072】以上の処理を全検索文書時間区分格納バッファ209に格納されている検索文書IDの全てに行う。処理後、検索文書単語頻度格納バッファ211に格納された情報は、全検索文書の単語情報として外部記憶装置4に格納、蓄積される（ステップC15）。

【0073】このようなデータベース作成処理により、検索対象となる各文書が分野別かつ時間別に分類され、その分類毎の単語出現頻度情報が外部記憶装置4に作成される。分野別の分類とは、内容的な要素による分類であり、図14に示す【大分類】（具体的には特許文献におけるIPCクラス）に当たる。時間別の分類とは、文書の作成期間あるは発行期間などを示す時間的な要素による分類であり、図14に示す【出願日】に当たる。

【0074】以下に、このような分類毎の単語出現頻度情報を用いた類似文書検索処理の動作について説明する。

【0075】（b）類似文書検索処理

図7は類似文書検索処理の動作を示すフローチャートである。

【0076】制御装置1は、まず、制御部1aの初期化部101を起動し、メモリ部1bの全バッファ部の初期化を行う（ステップD11）。

【0077】次に、検索文書大分類データ読み込み部103および検索文書時間区分データ読み込み部105を起動し、外部記憶装置4から検索文書大分類データファイルF2、検索文書時間区分データファイルF4を読み込んで検索文書大分類格納バッファ202、検索文書時間区分格納バッファ204にそれぞれ格納する。また、全検索文書大分類情報読み込み部114および全検索文書時間区分読み込み部115を起動し、外部記憶装置4から全検索文書大分類情報データファイルF5、全検索文書時間区分データファイルF6を読み込んで、全検索文書大分類情報格納バッファ208、全検索文書時間区分格納バッファ209にそれぞれ格納する（ステップD12）。

【0078】ここで、制御装置1は検索キー文書読み込み部119を起動し、入力装置2を通じてユーザが指定した検索キー文書の入力を受け付け、入力された検索キー文書のテキスト情報を検索キー文書格納バッファ212に格納する（ステップD13）。具体例として、例えば図26に示すような内容のテキスト文書が検索キー文

書の1つとして検索キー文書格納バッファ212に格納されたとする。

【0079】検索キー文書の読み込み後、制御装置1は検索キー文書格納バッファ212に格納された検索キー文書を先頭から順に走査していく(ステップD14)。その間、検索キー単語抽出部120を起動し、当該検索キー文書2に格納されたテキスト情報を形態素解析して単語の切り出しを行い、その切り出した各単語の中から文書内容を表すキー単語を抽出し、そのキー単語の単語種(例えば品詞情報)を検索キー文書単語情報格納バッファ213に格納する(ステップD15)。

【0080】続いて、検索キー文書単語出現頻度算出部121を起動し、当該検索キー文書全体での単語の出現頻度を算出し、その結果を検索キー文書単語情報格納バッファ213に格納する(ステップD16)。図27にこのときのデータ格納例を示す。この場合、単語と頻度が対について格納される。例えば、キー単語「写像」が当該検索キー文書の中で6回出現している場合は、頻度として「2」が格納される。

【0081】このようにして、検索キー文書の単語情報が得られると、制御装置1は、大分類毎かつ時間区分毎に検索文書と検索キーとの類似算出処理を行う。

【0082】すなわち、制御装置1は検索文書単語頻度読み込み部122を起動して、上記(a)のデータ作成処理によってデータベース外部記憶装置4上に作成された大分類かつ時間区分の検索文書単語出現頻度情報を1件ずつ読み込み(ステップD17、D18)、これを検索文書単語頻度格納バッファ211に順次格納していく(ステップD19)。

【0083】次に、ノルム情報算出部123を起動して検索キー文書単語情報格納バッファ213に格納された検索キー文書の単語頻度情報と、検索文書単語頻度格納バッファ211に格納された大分類別時間区分別検索文書の単語頻度情報に基づいて、検索キー文書に対する1次元ベクトルのノルム情報を算出し、その値を検索キー文書ノルム情報格納バッファ215に格納する(ステップD20)。図28にこのときのデータ格納例を示す。

【0084】ここで、ノルム情報について説明する。

【0085】ノルムとは、1次元のベクトルを $A = (a_1, a_2, \dots, a_n)$ としたとき、

$|A| = \sqrt{\text{(各ベクトル要素の二乗和)}}$

で表される値のことである。このノルム値は、後述する検索キー文書とデータベース内の検索文書との類似度計算で使用される。

【0086】図8のフローチャートを参照して、このノルム算出処理の具体的な説明を行う。ここでは、検索キー文書を対象とした場合でのノルム算出処理について説明する。

【0087】ノルム情報算出部123は、まず、検索キー文書単語情報格納バッファ213に格納された検索キ

ー文書の単語頻度値を“f”として、図32に示すように作業用変数バッファ219の単語頻度格納領域にセットする(ステップE12)。次に、検索文書単語頻度格納バッファ211に格納された検索文書の単語頻度値を“wg”として、作業用変数バッファ219の単語重み格納領域にセットする(ステップE13)。

【0088】ここで、“f”を“wg”で割った商の二乗を求め、その値を“nr”として、作業用変数バッファ219の作業用ノルム値格納領域に格納する(ステップE14)。この計算による処理は、検索キー文書の単語頻度に重みを付ける処理として実施する。つまり、検索キー文書の単語頻度は、大分類かつ時間区分における重要度として考えられ、頻度が多ければ、その単語の重要度は低く、頻度が少なければ、その単語の重要度は高くなる。したがって、“f”を“wg”で割った商とは、その単語の頻度に重みを付けた結果と言える。ここまでの処理を実施したら、次の単語に対する処理を行い(ステップE15)。

【0089】全ての単語について処理が終了した時点で(ステップE11のYes)、作業用変数バッファ219の作業用ノルム値格納領域に蓄積された“nr”の平方根を求め、その値を当該検索キー文書のノルム値として検索キー文書ノルム情報格納バッファ215に格納する(ステップE16)。

【0090】このような処理により、検索キー文書のノルム値が得られる。

【0091】次に、制御装置1は検索文書単語情報読み込み部116を起動し、外部記憶装置4に格納されている各検索文書の単語情報を大分類別に1つずつ読み込み、検索文書単語情報格納バッファ210に格納する(ステップD22)。

【0092】ここで、ノルム情報算出部123が検索文書単語情報格納バッファ210に格納された検索文書の単語情報と検索文書単語頻度格納バッファ211に格納された検索文書の単語頻度情報とを参照し、検索文書に対するノルム情報を算出し、その結果を検索文書ノルム情報格納バッファ214に格納する(ステップD23)。図24にこのときのデータ格納例を示す。

【0093】なお、この検索文書に対するノルム情報の具体的な算出処理については、参照する単語情報が検索文書単語情報という違いだけで、それ以外は上述した検索キー文書のノルム算出処理(図8)と同様であるため、ここではその説明を省略するものとする。

【0094】次に、制御装置1は共通単語抽出部124を起動して、検索キー文書単語情報格納バッファ213と検索文書単語情報格納バッファ210とで共通に格納されているキー単語を検出し、その共通単語を共通単語情報格納バッファ216に格納する(ステップD24)。

【0095】図27の検索キー文書単語情報格納バッファ213と図23の検索文書単語情報格納バッファ210の例では、両者に共通する単語として「文書」、「検索」が検出され、図29に示すように共通単語情報格納バッファ216に格納される。

【0096】次に、制御装置1は類似度算出部125を起動し、共通単語情報格納バッファ216に格納されている共通単語に基づき、検索キー文書と検索対象大分類の単語情報との類似度を所定の方法により算出し、その類似度を検索文書IDと対応付けて類似度格納バッファ217に格納する(ステップD25)。

【0097】本実施形態では、検索キー文書と検索対象との類似度をベクトル空間法により算出するものとする。

【0098】ベクトル空間法では、2つの一次元ベクトル、例えば、 $A = (a_1, a_2, \dots, a_n)$ 、 $B = (b_1, b_2, \dots, b_n)$ の類似度Sを次のように算出する。

$$【0099】S = A \cdot B / (|A| \cdot |B|)$$

分子の $A \cdot B$ は、2つの一次元ベクトルの内積であり、分母の $|A| \cdot |B|$ は、それぞれの一次元ベクトルのノルム(ベクトルの大きさ)の積である。この類似度Sは、 $0 \leq S \leq 1$ の範囲にあり、類似度が1に近いほど、2つの一次元ベクトルが類似していることになる。

【0100】ここで、単語情報を一次元ベクトルと考え、上記の例にあてはめると、検索キー文書の全ての単語の出現頻度がベクトルA、検索文書の全ての単語の出現頻度がベクトルBとなり、検索キー文書と検索文書との類似度はSという値で得られる。

【0101】この類似度算出処理について、図9を用いて詳しく説明する。

【0102】図9は上記ステップD25における類似度算出処理の動作を示すフローチャートである。

【0103】まず、類似度算出部125は、検索キー文書のノルム値を検索キー文書ノルム情報格納バッファ215から取得し、これを“na”として、図32に示す作業用変数バッファ219の検索キー文書ノルム格納領域にセットすると共に、検索文書のノルム値を検索文書ノルム情報格納バッファ214から取得し、これを“nb”として作業用変数バッファ219の検索文書ノルム格納領域にセットする(ステップF11)。

【0104】次に、類似度算出部125は、共通単語情報格納バッファ216に格納されている共通単語を参照し、その共通単語の数分だけ、以下のような処理を行う。

【0105】すなわち、検索キー文書単語情報格納バッファ213から検索キー文書における共通単語の頻度情報を取得し、その頻度値を“a”として作業用変数バッファ219の検索キー単語頻度格納領域にセットすると共に、検索文書単語情報格納バッファ210から検索文書における共通単語の頻度情報を取得し、その頻度値を

“b”として作業用変数バッファ219の検索対象単語頻度格納領域にそれぞれセットする(ステップF13)。

【0106】さらに、大分類別時間区分別の検索文書単語頻度格納バッファ211から上記共通単語の頻度情報を取得し、その頻度値を“w”として作業用変数バッファ219の作業用頻度格納領域にセットする(ステップF14)。

【0107】ここで、上記共通単語の内積を以下のようにして算出し、その値Rを作業用変数バッファ219の内積格納領域に累積する(ステップF15)。

$$【0108】R = (a/w) * (b/w)$$

ここまでの処理を共通単語情報格納バッファ216に格納されている共通単語の数分だけ行くと(ステップF12のYes)、類似度算出部125は、作業用変数バッファ219の各領域に格納された“R”、“na”、“nb”を用いて、以下のような演算を行って検索キー文書と検索文書との類似度Sを求める(ステップF16)。

$$【0109】S = R / (na * nb)$$

以上が類似度の算出処理である。

【0110】このようにして、検索キー文書と検索文書との類似度を大分類の時間区分の全ての検索文書について求める(ステップD21)。

【0111】この一連の処理を全ての時間区分数分について行くと(ステップD18のYes)、次の大分類に対する処理を行うべく、ステップD17に戻る。

【0112】全て大分類について上記同様の処理を行うと(ステップD17のYes)、算出類似度格納バッファ217には検索キー文書と全ての検索文書との類似度が格納されることになる。図30(a)にこのときのデータ格納例を示す。

【0113】ここで、制御装置1は類似度格納バッファ217に格納されている各検索文書に対する類似度値を高い順にソートして、検索結果出力バッファ218に格納する(ステップD26)。ソート後のデータ格納例を図30(b)に示す。そして、制御装置1は検索結果出力部126を起動し、検索結果出力バッファ218に格納された検索結果(ソート後のデータ)の内容を表示装置3に出力する(ステップD27)。検索結果は、例えば図31に示すような形態で出力するものとする。この例では、文書番号(ID)「2」、「1」、「3」…といった順で検索結果が出力されている。なお、検索結果の出力に際し、予め設定された閾値以上の類似度値を有する文書のみを対象として出力することで、検索結果として提示する文書数を制限するようにしても良い。

【0114】また、さらに別の検索キー文書があれば(ステップD28のNo)、ステップD13に戻って上記同様の処理を行うことになる。

【0115】このように、検索対象となる各文書を分野

別に分類すると共に、さらに、年別など時間的な区分で分類し、その分野別かつ時間区分別の単語出現頻度情報を作成することで、分野別の時間的要素に応じた各単語の重要度を得ることができる。これにより、検索キー文書と各検索文書との類似度計算を行う際に、検索キー文書と検索文書のそれぞれの単語頻度情報の中の単語毎の出現頻度を要素とする一次元ベクトルにおいて、各単語の出現頻度に対し、分野別の時間的要素に応じた各単語の重要度を加味して、精度の高い類似検索を実現することができる。

【0116】なお、上記実施形態では、検索キーとして与えられた文書の分野、時間区分を全分野全時間区分の検索文書とのマッチングにより判断したが、検索キー文書の分野あるいは時間区分が明らかである場合には、その分野、時間区分に限定したマッチングを行うといった方法を探っても良い。

【0117】また、上記実施形態では、各単語の出現頻度分野毎の時間的な要素に応じた重み付けを行うようにしたが、少なくとも時間的な要素に応じた単語の重み付けを行うことでも良い。ただし、上記図1で説明したように、特徴単語の時間的な変化は分野によって異なるため、多種の分野に亘って文書が存在する場合には、上記実施形態のように分野と時間の両方の要素を加味して単語の重み付けを行うことが望ましい。

【0118】また、本発明の類似文書検索装置は、例えば特許分野における引用文献の検索の他、一般的なパーソナルコンピュータ等におけるファイル管理など、検索を必要とする技術一般に広く適用できるものである。

【0119】また、上述した実施形態において記載した手法は、コンピュータに実行させることのできるプログラムとして、例えば磁気ディスク（フロッピーディスク、ハードディスク等）、光ディスク（CD-ROM、DVD等）、半導体メモリなどの記録媒体に書き込んで各種装置に適用したり、通信媒体により伝送して各種装置に適用することも可能である。本装置を実現するコンピュータは、記録媒体に記録されたプログラムを読み込み、このプログラムによって動作が制御されることにより、上述した処理を実行する。

【0120】

【発明の効果】以上詳記したように本発明によれば、各文書毎の単語の出現頻度に各文書の時間的な要素に応じた重み付けを行うようにしたため、時間別の単語の重要度を得ることができ、各検索対象文書と検索キー文書との類似度を求める際にその単語重要度を加味することで、高精度の類似検索を実現することができる。

【0121】また、各文書を分野別に分類し、その分野別の時間的な要素に応じた重み付けを行うことで、分野によって異なる特徴単語の時間的な変化を反映させた単語の重要度を得ることができる。このような分野別かつ時間別の単語の重要度を各検索対象文書と検索キー文書

との類似度計算に加味することで、さらに高精度の類似検索を実現することができる。

【図面の簡単な説明】

【図1】本発明の類似文書検索装置の概要を説明するための図。

【図2】本発明の一実施形態に係る類似文書検索装置の構成を示すブロック図。

【図3】図1に示す制御装置の内部構成を示すブロック図。

【図4】データベース作成処理の動作を示すフローチャート。

【図5】上記図4のステップA11における検索文書登録処理の具体的な動作を示すフローチャート。

【図6】上記図4のステップA12における検索文書単語出現頻度作成処理の具体的な動作を示すフローチャート。

【図7】類似文書検索処理の動作を示すフローチャート。

【図8】上記図7のステップD20におけるノルム算出処理の具体的な動作を示すフローチャート。

【図9】上記図7のステップD25における類似度算出処理の具体的な動作を示すフローチャート。

【図10】検索文書大分類識別子データファイル（F1）の一例を示す図。

【図11】検索文書大分類データファイル（F2）の一例を示す図。

【図12】検索文書時間区分識別子データファイル（F3）の一例を示す図。

【図13】検索文書時間区分データファイル（F4）の一例を示す図。

【図14】検索文書格納バッファの内容を示す図。

【図15】検索文書大分類識別子データ格納バッファの内容を示す図。

【図16】検索文書大分類格納バッファの内容を示す図。

【図17】検索文書時間区分識別子データ格納バッファの内容を示す図。

【図18】検索文書時間区分データ格納バッファの内容を示す図。

【図19】検索文書大分類情報格納バッファの内容を示す図。

【図20】全検索文書大分類情報格納バッファの内容を示す図。

【図21】検索文書時間区分格納バッファの内容を示す図。

【図22】全検索文書時間区分格納バッファの内容を示す図。

【図23】検索文書単語情報格納バッファの内容を示す図。

【図24】検索文書ノルム情報格納バッファの内容を示す図。

す図。

【図25】検索文書単語頻度格納バッファの内容を示す図。

【図26】検索キー文書格納バッファの内容を示す図。

【図27】検索キー文書単語情報格納バッファの内容を示す図。

【図28】検索キー文書ノルム情報格納バッファの内容を示す図。

【図29】共通単語情報格納バッファの内容を示す図。

【図30】類似度格納バッファの内容（ソート前とソート後）を示す図。

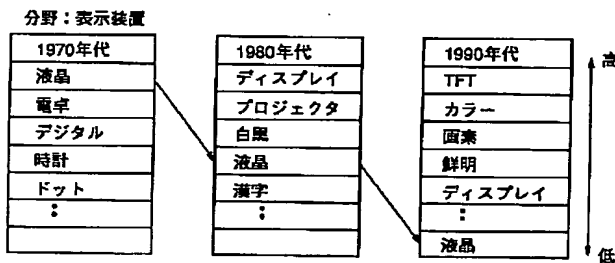
【図31】検索結果出力バッファの内容を示す図。

【図32】作業用変数バッファの内容を示す図。

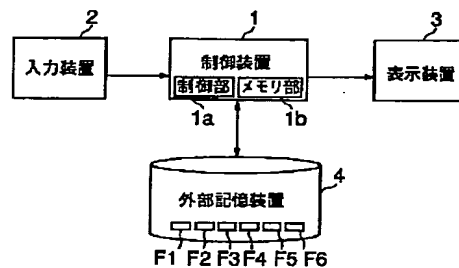
【符号の説明】

- 1…制御装置
- 1a…制御部
- 1b…メモリ部
- 2…入力装置
- 3…表示装置
- 4…外部記憶装置
- F1…検索文書大分類識別子データファイル
- F2…検索文書大分類データファイル
- F3…検索文書時間区分識別子データファイル
- F4…検索文書時間区分データファイル
- F5…全検索文書大分類情報データファイル
- F6…全検索文書時間区分データファイル

【図1】



【図2】



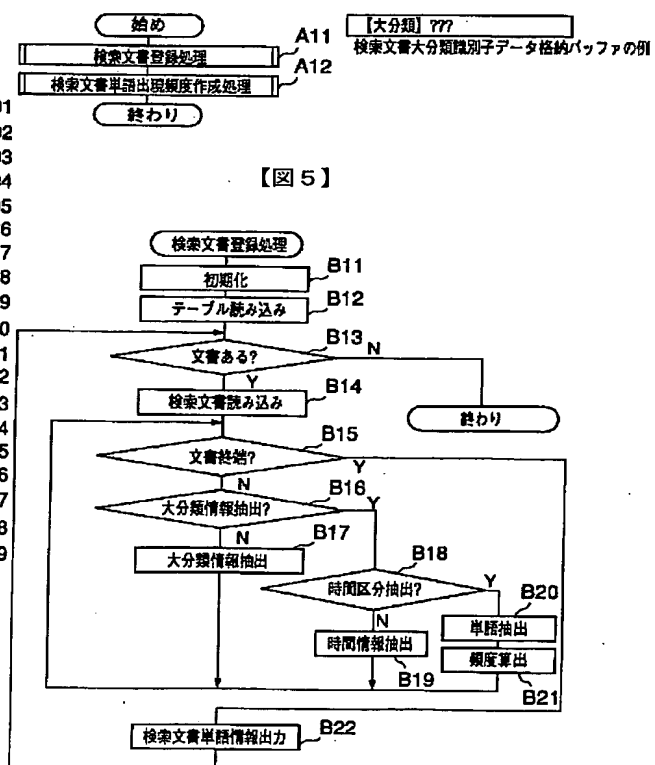
【図3】

【図4】

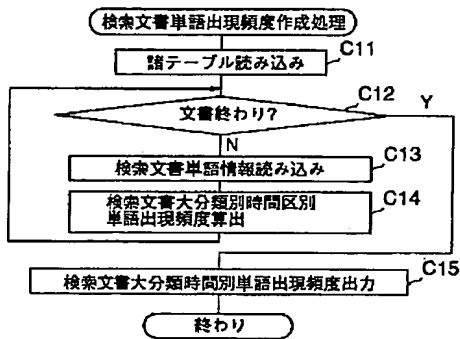
【図15】



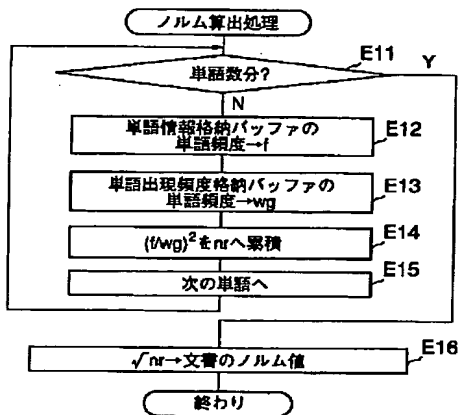
【図5】



【図6】



【図8】



【図10】

検索文書大分類別子データファイルの内容

【大分類】???

【図12】

検索文書時間区分別子データファイルの内容

【出願日】???年

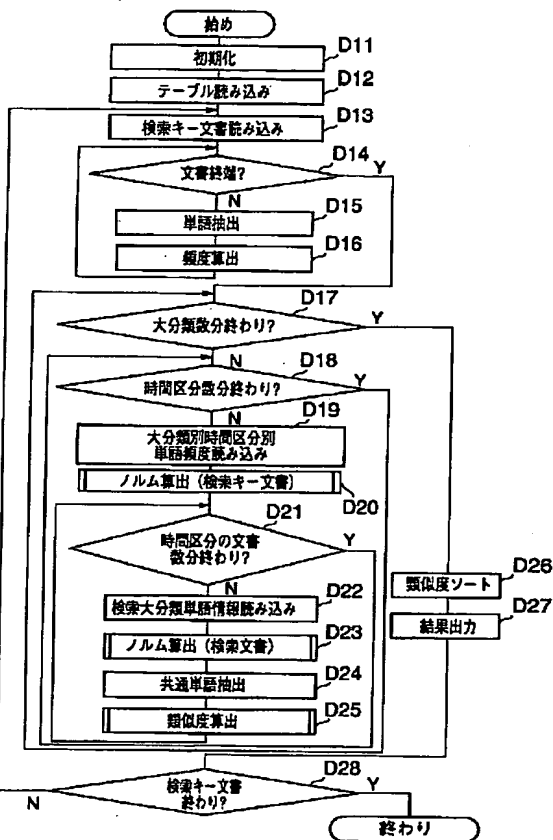
【図13】

検索文書時間区分別子データファイルの内容
(時間区分ID、時間区分)1: 1999
2: 1998
3: 1997
...

【図17】

【出願日】*年
検索文書時間区分別子データ格納バッファの例

【図7】



【図9】

【図11】

検索文書大分類データファイルの内容
(大分類ID,大分類)1: A01
2: A21
3: A22
...
118: H05

【図16】

大分類ID	大分類データ
1	A01
2	A21
3	A22
...	...
118	H05

検索文書大分類格納バッファの例

【図19】

大分類ID	大分類情報
1	A01

検索文書大分類情報格納バッファの例1

【図14】

【大分類】A01
【出願日】1999年(平成5年)9月27日
【発明名称】
類似文書検索装置および類似文書検索方法
【発明の属する技術分野】
<一般的技術分野>
文書検索技術
<特定技術>
特徴単語抽出技術、類似文書検索技術、
【従来の技術】
近年、大量の電子化された文書データが流通するようになり、自動分類等を行う目的で、
【発明の解決しようとする課題】
類似文書検索において、各文書から特徴単語の抽出する際、その文書の特徴づける単語を抽出することが非常に

検索文書格納バッファの例

【図18】

時間区分ID	時間区分
1	1999
2	1998
3	1997
⋮	⋮

検索文書時間区分格納バッファの例

【図20】

検索文書ID	大分類ID
1	1
2	21
3	5
4	3
⋮	⋮

全検索文書大分類情報格納バッファの例

【図23】

単語	頻度
文書	2
類似	5
技術	1
⋮	⋮
検索	7

検索文書単語情報格納バッファの例

【図24】

31,341
検索文書ノルム情報格納バッファの例

【図28】

【図21】

時間区分ID	時間区分情報
1	1999

検索文書時間区分格納バッファの例1

【図22】

検索文書ID	時間区分ID
1	1
2	1
3	3
4	2
⋮	⋮

全検索文書時間区分格納バッファの例

【図27】

単語	頻度
文書	3
キーワード	2
写像	6
⋮	⋮
検索	10

検索キーワード単語情報格納バッファの例

14,484
検索キーワードノルム情報格納バッファ

【図31】

類似文書検索結果
<文書番号>
2
1
3
⋮

検索結果出力バッファの例

【図25】

大分類ID: 1
時間区分ID: 1
単語 頻度
文書 23
類似 6
表示 120
⋮
区分 9

時間区分ID: 2
単語 頻度
類似 44
表示 63
抽出 2
⋮
課題 89

時間区分ID: N
単語 頻度
計算 44
類似 161
表示 33
⋮
検索 83

大分類ID: 2
時間区分ID: 1
単語 頻度
溶解 48
水消 66
熱 12
⋮
分解 1

時間区分ID: 2
単語 頻度
使用 3
課題 231
破壊 41
⋮
塩素 88

時間区分ID: N
単語 頻度
塩素 160
自動 31
生成 71
⋮
監視 9

【発明の名称】
文書検索装置
【発明の属する技術分野】
<一般的技術分野>
文書検索技術
<特定技術>
特徴単語抽出技術、文書検索技術
【従来の技術】
フィールド検索は一般的に活用されているが、その検索方法は、各フィールド毎に検索キーワードの成立・不成立をチェックするものである。しかし、対象となる文書
⋮
【発明が解決しようとする課題】
様々な文書の中には、一つ文書において、その構造がいくつかのフィールドに分かれているものがある。そうした文書では、より多くのフィールドに分布している単語がその文書の内容を表わすことが多い。キーワードを

検索キーワード格納バッファの例

【図26】

大分類ID: M
時間区分ID: 1
単語 頻度
基礎 48
木材 66
時間 12
⋮
設計 1

検索文書単語頻度格納バッファの例

時間区分ID: 2
単語 頻度
監視 9
時間 24
静止 13
⋮
自動 6

時間区分ID: N
単語 頻度
使用 201
自動 21
脱落 81
⋮
固定 65

【図30】

検索文書ID	類似度
1	0.02165
2	0.03252
3	0.00145
⋮	⋮

類似度格納バッファの例(ソート前)

大分類ID	類似度
2	0.03252
1	0.02165
3	0.00145
⋮	⋮

類似度格納バッファの例(ソート後)

【図29】

単語
文書
検索
⋮

共通単語情報格納バッファの例

【図32】

検索文書ID	1
単語頻度:f	12
単語重み:wg	7
作業用ノルム値:nr	23.448
検索キー文書ノルム:na	49.223
検索対象ノルム:nb	76.331
検索キー単語頻度:a	34
検索対象単語頻度:b	8
作業用頻度:w	11
内積:n	76
類似度:s	0.0384

作業用変数バッファの例

フロントページの続き

(72) 発明者 中里 茂美
東京都青梅市末広町2丁目9番地 株式会
社東芝青梅工場内

(72) 発明者 齋藤 裕美
東京都青梅市末広町2丁目9番地 株式会
社東芝青梅工場内

(72) 発明者 仁科 卓哉
東京都青梅市新町3丁目3番地の1 東芝
コンピュータエンジニアリング株式会社内

(72) 発明者 中本 幸夫
東京都青梅市新町3丁目3番地の1 東芝
コンピュータエンジニアリング株式会社内

(72) 発明者 山崎 弘
東京都青梅市新町3丁目3番地の1 東芝
コンピュータエンジニアリング株式会社内

(72) 発明者 松隈 剛
東京都青梅市新町3丁目3番地の1 東芝
コンピュータエンジニアリング株式会社内

Fターム(参考) 5B075 ND03 NK02 NK10 NR05 NR12
PR04 PR06 PR10 UU05